International Journal of Education and Teaching Zone
Volume 5, Issue 1, 2026, pp. 217-236

https://doi.org/10.57092/ijetz.v5i1.546
p-ISSN: 2963-7899 | e-ISSN: 2830-7925

Original Article

Development and Validation of a Critical Thinking Skills Instrument for Grade 11 Biology Students Based on Ennis's Framework

Try Susanti[1], Dwi Gusfarenie[1], Nanda Gusriani[1]*, Diandara Oryza[1], Nining Nuraida[1]

Abstract: The ability to think critically is a key competency required of students in the 21st century, particularly in science education. This study aimed to develop and empirically evaluate a critical thinking skills assessment instrument for Grade 11 senior high school students in biology learning. The instrument was developed using the ADDIE development model, encompassing the analysis, design, development, implementation, and evaluation stages. Empirical testing was conducted during the implementation stage with Grade 11 senior high school students in Jambi City. Item analysis was performed to examine content validity, empirical validity, reliability, item difficulty, and discrimination indices. Empirical data were analyzed using ANATES version 4.0.9. The results indicated that 22 out of 56 items met the predefined validation criteria, demonstrating acceptable levels of item difficulty and discrimination power. The reliability coefficient of the finalized instrument was 0.68, which is considered acceptable for an exploratory educational assessment instrument. These findings suggest that the developed instrument possesses adequate psychometric properties for measuring students' critical thinking skills in biology learning and can support future instructional evaluation and research within similar educational contexts.

Keywords :
Biology education; Critical thinking skills; Instrument development; Item analysis

Author Affiliation:
[1] Biology Education Study Program, UIN Sulthan Thaha Saifuddin, Jambi, Indonesia

***Corresponding author(s):**
Nanda Gusriani, UIN Sulthan Thaha Saifuddin Jambi, Jambi, Indonesia
nanda.gusriani@uinjambi.ac.id

Article History:
Received 1 November 2025; Revised 16 January 2026; Accepted 14 February 2026
Available online 28 February 2026

INTRODUCTION

Biology continues to develop rapidly in the era of the Industrial Revolution 4.0, particularly through advances in genetic editing, neurotechnology, and biotechnology that influence various sectors, including education (Schwab, 2017). These developments require students not only to master scientific knowledge but also to develop higher-order thinking skills to interpret data, evaluate evidence, and make reasoned decisions

(Krathwohl, 2002). Within the framework of 21st-century education, schools are expected to prepare high-quality individuals capable of adapting to complex global challenges Partnership for 21st Century Learning. Four core competencies are emphasized: critical thinking or problem solving, creativity, communication, and collaboration (World Economic Forum, 2020). In a study, many results were obtained. Among them, critical thinking is considered a fundamental competency that supports reflective judgment and responsible decision making (Ennis, 2018; Facione, 2020), which involves analyzing alternatives, evaluating arguments, integrating information, and formulating reasonable conclusions (Halpern, 2014; Willingham, 2019).

In Indonesia, the importance of critical thinking is reinforced through Ministerial Regulation No. 20 of 2016 on Graduate Competency Standards, which explicitly identifies critical thinking as a core competency for senior high school graduates. National examination items have increasingly incorporated Higher Order Thinking Skills (HOTS) . Research on HOTS-based assessment in biology education has shown that valid and reliable instruments are essential for measuring students' critical thinking abilities (Rahmi et al., 2021; Ulfa & Kuswanti, 2020). However, the inclusion of HOTS-type items in large-scale assessments does not automatically ensure comprehensive measurement of the multidimensional construct of critical thinking, as studies on Indonesian national examinations indicate that HOTS items are often limited to the "Analyze" level and lack variation in cognitive process dimensions. Research on domain-specific instrument development demonstrates the importance of systematic construct validation and psychometric testing to ensure reliability and validity across student populations (Abidin & Fatimatuzzaro, 2025; Fajar & Suryani, 2023).

In this study, critical thinking is conceptualized based on Ennis's framework, which defines it as reasonable and reflective thinking focused on deciding what to believe or do (Ennis, 2018; Paul, 2018), and operationalizes it into measurable abilities such as analyzing arguments, evaluating evidence, drawing inferences, and making justified decisions. HOTS principles are integrated with this framework to guide item construction, ensuring that higher-order tasks are clearly grounded in defined critical thinking dimensions rather than loosely categorized advanced content recall (Khoeriyah *et al*., 2025). Instrument quality is evaluated through evidence of construct validity, empirical validity, reliability, item difficulty, and discrimination indices (Rahmi *et al*., 2021), enabling accurate differentiation of students' reasoning levels and diagnostic feedback for instructional improvement (Yokhebed *et al*., 2025).

Despite the growing emphasis on critical thinking in the Indonesian biology curriculum and the widespread integration of Higher Order Thinking Skills (HOTS) in national assessments, a significant gap remains in the availability of empirically validated, subject-specific instruments to measure this construct. Existing classroom assessments in biology often rely on general critical thinking frameworks or focus predominantly on lower-order cognitive tasks, failing to capture the multidimensional nature of critical thinking as operationalized within biological contexts (Abidin & Fatimatuzzaro, 2025; Yokhebed *et al*., 2025). Furthermore, while frameworks such as Ennis's model of critical thinking have been widely adopted theoretically, their systematic operationalization into a

valid and reliable biology-specific assessment tool for senior high school students remains underexplored.

To address this gap, the present study develops and empirically analyzes a biology-specific critical thinking assessment instrument for Grade 11 students using systematic item validation procedures. Grounded in Ennis's conceptual framework and aligned with HOTS principles, this instrument is designed to measure key critical thinking dimensions including analyzing arguments, evaluating evidence, drawing inferences, and making justified decisions within the context of biology learning. By integrating theoretical grounding with rigorous psychometric analysis, this research seeks to provide a validated tool that supports evidence-based evaluation of students' critical thinking skills. This study, entitled "Development and Validation of a Critical Thinking Skills Instrument for Grade 11 Biology Students Based on Ennis's Framework," contributes to the advancement of assessment practices in biology education by offering an instrument that is both contextually relevant and psychometrically sound.

METHOD

This study employed the ADDIE (Analysis, Design, Development, Implementation, and Evaluation) development model as a systematic framework for designing and validating the critical thinking assessment instrument (Branch, 2009). The analysis phase involved a preliminary needs assessment through classroom observations and informal interviews with biology teachers to identify challenges in assessing students' critical thinking skills. The findings indicated that most existing assessments primarily measured factual recall and conceptual understanding rather than higher-order reasoning, a finding consistent with previous research on biology classroom assessments (Septiany *et al*., 2024). Curriculum documents and Grade 11 biology content standards were also analyzed to ensure alignment between the instrument content and learning objectives.

The design phase, a test blueprint was constructed to operationalize critical thinking based on Ennis's framework (Ennis, 1991; Ennis, 2018), which includes five main indicators and twelve sub-indicators as elaborated in his streamlined conception of critical thinking. Each test item was explicitly mapped to a specific Ennis indicator, a relevant biology topic, and a cognitive process requiring analysis, evaluation, inference, or justification. To ensure that the instrument measured critical thinking rather than recall, items were developed in contextual and problem-based formats requiring students to interpret data, evaluate arguments, explain reasoning, and justify decisions. Essay-type questions were selected to allow students to demonstrate structured and evidence-based reasoning, an approach supported by recent research on critical thinking assessment in science education (Dini & Kuswanto, 2025; Ariyanti & Rahayu, 2025).

During the development phase, the draft instrument was validated by three purposively selected experts consisting of an educational evaluation expert, a biology education lecturer, and a senior biology teacher with more than ten years of teaching experience. The experts assessed construct alignment, content relevance, clarity of language, contextual accuracy, and consistency with critical thinking indicators. In this

research, an assessment score is of course needed, the validity score of which is calculated using the formula:

$$V=TSe/TSh \times 100\% \quad (1)$$

where TSe represents the total empirical score and TSh represents the maximum expected score (A'yun *et al.*, 2022; Retnawati, 2016). Validity categories were interpreted based on established benchmarks as presented in Table 1, adapted from Akbar (2013). In addition to quantitative scoring, qualitative feedback from validators was systematically documented and integrated into revisions, particularly to improve item clarity, strengthen alignment with Ennis's indicators, and eliminate ambiguity (Septiany *et al.*, 2024). Inter-rater agreement was examined by comparing scoring consistency among validators prior to calculating the average validity score to enhance transparency of construct validation (O'Connor & Joffe, 2020; Syafril *et al.*, 2021).

Where V is the validity percentage, TSe represents the total empirical score obtained from expert validators, and TSh is the maximum expected score (A'yun *et al.*, 2022; Retnawati, 2016). The criteria for interpreting the validation scores are presented in Table 1, adapted from Akbar (2013).

Table 1. Validity Category

No.	Percentage of Assessment Scores (%)	Category
1.	$85{,}00 \leq V \leq 100$	Very Valid
2.	$70{,}00 \leq V < 85{,}00$	Valid
3.	$50{,}00 \leq V < 70{,}00$	Less valid
4.	$00{,}00 \leq V < 50{,}00$	Invalid

The implementation phase involved administering the revised instrument to 103 Grade 11 students at MA Negeri 2 Jambi City. Participants were selected using cluster sampling from three science-track classes and were aged between 16 and 17 years. The sample size met common classical test theory recommendations, which suggest a minimum of 5–10 respondents per item to produce stable item statistics (DeVellis, 2017; Nunnally & Bernstein, 1994). Prior to testing, students had completed the relevant biology topics assessed in the instrument to ensure content familiarity and readiness. Data collection was conducted during regular instructional hours under standardized administration procedures to minimize testing bias. Students were informed about the purpose of the study and assured that their responses would remain confidential and would not affect their academic grades. The instrument was administered in a paper-based format and completed within approximately 45–60 minutes. After collection, responses were coded and analyzed using classical test theory procedures, including item difficulty index, discrimination index, and internal consistency reliability analysis (Cronbach's alpha). These analyses were undertaken to evaluate the psychometric quality of each item and to determine whether further revision or elimination was necessary before finalizing the instrument for broader application.

Item analysis was conducted using ANATES version 4.0.9 software, a widely used application for educational measurement in Indonesian contexts (Karnengsih *et al*., 2021). Empirical validity was examined through item-total correlation coefficients to determine the consistency between each item and the overall test score. Internal consistency reliability was calculated using Cronbach's Alpha, with interpretation criteria adapted from Erfan *et al*. (2020) as shown in Table 2. Item difficulty indices were analyzed to classify items as easy, moderate, or difficult, while discrimination indices were calculated to evaluate each item's ability to differentiate between high- and low-performing students (Arikunto, 2015; Retnawati, 2016). Interpretation of difficulty and discrimination scores following standard educational measurements is presented in Table 3.

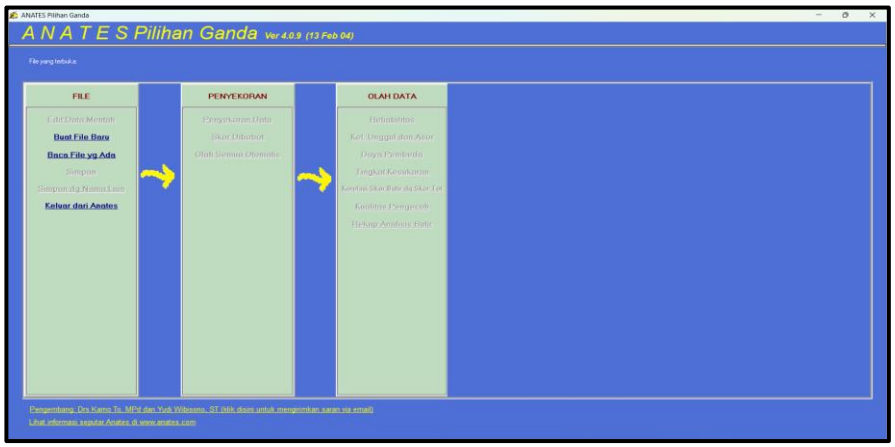

Figure 1. Initial Display of The ANATES Version 4.0.9 Software

In the evaluation phase, item analysis was conducted using ANATES version 4.0.9 software based on classical test theory principles. Empirical validity was examined through item-total correlation coefficients to determine the consistency between each item and the overall test score. Internal consistency reliability was calculated using Cronbach's Alpha. Item difficulty indices were analyzed to classify items as easy, moderate, or difficult, while discrimination indices were calculated to evaluate each item's ability to differentiate between high and low performing students. The interpretation of reliability, difficulty, and discrimination values followed commonly accepted educational measurement standards. Through this structured development and empirical validation process, the instrument was refined to ensure both theoretical alignment with critical thinking constructs and adequate psychometric quality for use in senior high school biology assessment. The test items are considered reliable if they meet the instrument reliability correlation coefficient criteria shown in Table 2 (Erfan *et al*., 2020).

Table 2. Instrument Reliability Correlation Coefficient Criteria

No.	Correlation Coefficient	Correlation	Interpretation/ Reliablity
1.	$0,90 \leq r \leq 1,00$	Very High	Very stable/very good
2.	$0,70 \leq r < 0,90$	High	Stable/good
3.	$0,40 \leq r < 0,70$	Moderate	Fairly stable/fairly good
4.	$0,20 \leq r < 0,40$	Low	Unstable/poor
5.	$r < 0,20$	Very Low	Very unstable/very poor

Table 2 presents the criteria for interpreting instrument reliability based on the magnitude of the correlation coefficient (r). A coefficient between 0.90 and 1.00 indicates very high reliability, meaning the instrument demonstrates excellent internal consistency and produces highly stable and dependable results. Coefficients ranging from 0.70 to 0.89 are categorized as high, reflecting good stability and acceptable reliability for most educational research purposes. Values between 0.40 and 0.69 represent moderate reliability, suggesting that the instrument is fairly consistent but may still require refinement to improve precision. A coefficient between 0.20 and 0.39 is considered low, indicating limited stability and potential weaknesses in item consistency. Finally, a coefficient below 0.20 reflects very low reliability, meaning the instrument is highly unstable and not suitable for valid measurement without substantial revision.

RESULT AND DISCUSSION

Analysis Phase

During the analysis phase, a structured needs analysis and curriculum analysis were conducted to provide an empirical basis for instrument development. The needs analysis involved semi-structured interviews with three Grade 11 biology teachers at MA Negeri 2 Jambi City and classroom observations conducted over two learning sessions. The interview protocol focused on current assessment practices, the integration of Higher Order Thinking Skills (HOTS) in test construction, and challenges in measuring students' critical thinking skills. Interview data were transcribed and coded thematically to identify recurring patterns, following established qualitative analysis procedures (Braun & Clarke, 2021; Fereday & Muir-Cochrane, 2006). The findings indicated that although teachers recognized the importance of critical thinking, most assessment instruments predominantly emphasized factual recall and short-answer conceptual questions, a gap also identified in previous research on biology classroom assessments (Septiany *et al*., 2024; Sintia & Yuliani, 2024). Teachers also reported limited access to validated subject-specific instruments designed explicitly to measure critical thinking in biology (Rahmi *et al*., 2021; Abidin & Fatimatuzzaro, 2025). These findings provided empirical justification for developing a structured and psychometrically tested instrument.

The curriculum analysis examined official Merdeka Curriculum documents, including competency standards, learning objectives, and assessment guidelines for Grade 11 biology (Kemendikbudristek, 2022a, 2022b). Specific curriculum components emphasizing analysis, evaluation, reasoning, and problem-solving were systematically mapped against Ennis's critical thinking indicators (Ennis, 1991, 2018). This mapping process ensured that each proposed test indicator aligned with both curriculum expectations and theoretical constructs of critical thinking. For example, curriculum outcomes requiring students to analyze ecosystem interactions or evaluate environmental issues were linked to Ennis's indicators of inference, argument analysis, and strategic decision-making (Ratna *et al*., 2025; Santosa *et al*., 2023). This alignment guided the development of the test blueprint and ensured content validity. Recent studies on Merdeka Curriculum implementation confirm that project-based and inquiry-based learning

approaches are designed to strengthen 21st-century skills, including critical thinking (Sari *et al*., 2024; Wahyuningtyas *et al*., 2025).

Although the Merdeka Curriculum promotes student-centered, contextual, and project-based learning aimed at strengthening higher-order thinking skills (Kemendikbudristek, 2022a; Pamelia & Hariani, 2021), the needs analysis revealed inconsistencies between curricular expectations and assessment practices at the classroom level. This finding aligns with research indicating that while curriculum reforms emphasize HOTS, assessment instruments often lag behind in terms of validation and contextual relevance (A'yun et al., 2022; Fadlilah & Indana, 2025). Therefore, the analysis phase not only established the contextual relevance of critical thinking assessment within the Merdeka Curriculum framework but also identified a measurable gap between intended competencies and existing evaluation instruments (Almunawarah *et al*., 2023; Septiany *et al*., 2024). These findings directly informed the operationalization of critical thinking indicators into biology-specific test items, ensuring that the instrument was both theoretically grounded and contextually responsive.

Design Phase

The product design phase constitutes a critical foundation in research and development, as it systematically translates pedagogical standards into measurable assessment components (Suraida *et al*., 2025; Branch, 2009). In this study, the design phase began with the development of a detailed test blueprint that aligned Grade 11 Biology learning outcomes in the Merdeka Curriculum with Ennis's five critical thinking aspects and twelve sub-indicators (Ennis, 1991, 2018; Kemendikbudristek, 2022). A matrix alignment procedure was employed to map curriculum competencies (*Capaian Pembelajaran*) to specific critical thinking processes such as argument analysis, inference, evaluation, decision-making, and strategic reasoning (Ratna *et al*., 2025; Santosa *et al*., 2023). This mapping ensured content validity and conceptual coherence between curriculum demands and the theoretical construct of critical thinking (Retnawati, 2016; Haynes *et al*., 1995).

Based on this blueprint, an initial pool of 15 essay-type items was drafted. Each item was explicitly coded according to (1) the targeted Ennis indicator, (2) the relevant Biology topic, and (3) the intended cognitive operation (e.g., analyzing experimental data, evaluating biological claims, constructing justified conclusions). The selected content domains cells, tissues, and human organ systems were chosen because they represent foundational yet conceptually complex topics that require integrative reasoning, interpretation of biological processes, and contextual application (Septiany *et al*., 2024; Sintia & Yuliani, 2024). These topics were considered suitable for operationalizing all five Ennis aspects, including providing simple explanations, building basic skills, making inferences, giving further explanations, and setting strategies and tactics (Ennis, 1991; Almunawarah *et al*., 2023).

To promote higher-order thinking and avoid recall-based assessment, items were constructed using contextual scenarios, observational data, diagrams, and real-life biological problems that required students to justify reasoning and evaluate evidence (Yu

& Zin, 2023; Dini & Kuswanto, 2025). Each question was reviewed to ensure clarity of wording, cognitive demand at the analysis–evaluation level of Bloom's revised taxonomy (Krathwohl, 2002), and contextual relevance to students' learning experiences (Ariyanti & Rahayu, 2025). This internal review process functioned as a preliminary quality control step prior to formal expert validation (Cohen *et al*., 2018).

Based on this blueprint, an initial pool of 15 essay-type items was drafted. Each item was explicitly coded according to (1) the targeted Ennis indicator, (2) the relevant Biology topic, and (3) the intended cognitive operation (e.g., analyzing experimental data, evaluating biological claims, constructing justified conclusions). The selected content domains cells, tissues, and human organ systems were chosen because they represent foundational yet conceptually complex topics that require integrative reasoning, interpretation of biological processes, and contextual application. These topics were considered suitable for operationalizing all five Ennis aspects, including providing simple explanations, building basic skills, making inferences, giving further explanations, and setting strategies and tactics.

Answer keys and analytic scoring rubrics were developed concurrently with item construction to maintain alignment between the intended critical thinking processes and assessment criteria. The rubrics specified performance descriptors for each indicator, thereby facilitating objective and transparent scoring in accordance with the assessment principles of the Merdeka Curriculum (Kemendikbudristek, 2022; Pamelia & Hariani, 2021). Through this structured procedure blueprint construction, curriculum–construct mapping, item drafting, cognitive verification, and rubric development the instrument was systematically prepared for expert validation in the subsequent development phase (Abidin & Fatimatuzzaro, 2025; A'yun *et al*., 2022).

Development Phase

This study produced a biology-based test instrument designed to measure students' critical thinking skills. The initial validation process involved two subject-matter experts: a senior biology teacher at MA Negeri 2 Jambi City (Validator A) and a biology education lecturer at Bengkulu State University (Validator B). Both validators were selected purposively based on their professional experience in assessment design and biology instruction. Although the number of validators was limited, the selection ensured representation from both practitioner and academic perspectives to examine content relevance, construct alignment, clarity of language, and contextual appropriateness. A summary of the expert validity results of the test instrument is shown in Table 3.

Table 3. Recapitulation of the Expert Validity of the Question Instrument

Validator	Percentage	Criteria
A	78,47	Valid
B	79,86	Valid
Average	79,17	Valid

The results of expert validation indicated percentage scores of 78.47% (Validator A) and 79.86% (Validator B), with an overall average of 79.17%, categorized as "valid." These findings suggest that the instrument demonstrated satisfactory alignment with critical thinking dimensions in terms of material accuracy, construction quality, and linguistic clarity. The involvement of expert review is consistent with recommendations that instrument development requires systematic validation procedures to enhance content validity and measurement quality.

To enhance transparency, inter-rater consistency between two validators was examined by comparing scoring patterns before averaging validity scores (O'Connor & Joffe, 2020; Syafril *et al*., 2021). A percentage difference below 2% indicated acceptable agreement (McHugh, 2012). Qualitative feedback was categorized into three areas: technical presentation, linguistic refinement, and construct strengthening. Revisions addressed ambiguous wording, sentence length, visual clarity, and alignment with Ennis's critical thinking dimensions, and were cross-checked against the original blueprint to maintain construct validity (Ennis, 2018; Septiany *et al*., 2024; Retnawati, 2016).

Following theoretical validation, empirical testing was conducted to examine item performance in real classroom settings (DeVellis, 2017). The revised instrument was administered to Grade 11 students under standardized conditions (Cohen *et al*., 2018). Analysis included item-total correlations, reliability, difficulty index, and discrimination power to assess item effectiveness in differentiating critical thinking levels (Arikunto, 2015; Retnawati, 2016). This two-stage validation process strengthened both content validity and psychometric robustness (Messick, 1995; Haynes *et al*., 1995). Strong content validity was confirmed, supported by quantitative ratings and qualitative feedback from experts, consistent with prior findings (A'yun *et al*., 2022).

Table 4. Comments and Suggestions on Expert Validation Results

No.	Comments and Suggestions
1.	The font size on the image is very messy and a bit hard to read.
2.	The use of words is good, but the sentence structure still needs to be improved by using effective sentences.
3.	If the introductory sentence of the statement/question is less relevant, respondents will likely have difficulty understanding the question. Improve the question narrative.
4.	The image size is enlarged so that the text/image caption can be read.
5.	The image is correct, but needs to be made clearer.
6.	The critical thinking dimensions are in accordance with the test items, but improve the introductory and supporting sentences of the questions.
7.	There are several errors in writing the introductory sentence of the question so that it has a double meaning.
8.	Fix the test items that have too long sentences.

Table 4 presents the qualitative feedback from expert validation. Although the items were considered conceptually aligned with critical thinking dimensions, several technical and linguistic revisions were recommended. Feedback primarily addressed improving visual clarity (e.g., font and image size, image resolution), refining sentence structure for conciseness and effectiveness, and eliminating ambiguous introductory statements.

Experts also highlighted the need to strengthen question narratives and reduce overly long sentences to improve readability and minimize cognitive load. These revisions were necessary to enhance clarity, prevent misinterpretation, and strengthen the instrument's validity and reliability prior to large-scale implementation.

Implementation Phase

The implementation phase involved administering the finalized instrument under controlled classroom conditions to evaluate its empirical performance (Cohen *et al*., 2018; Fraenkel *et al*., 2012). The field trial was conducted during regular Biology class sessions with 103 Grade 11 science-track students at MA Negeri 2 Jambi City, aged 16–17 years. The participants represented three intact classes selected through cluster sampling (Creswell & Creswell, 2018). All students had completed the Biology topics assessed in the instrument prior to testing to ensure content familiarity and readiness (Septiany et al., 2024). The sample size of 103 students met classical test theory recommendations for stable estimation of item parameters, with a minimum of 5–10 respondents per item considered adequate for item analysis (DeVellis, 2017; Nunnally & Bernstein, 1994).

The instrument comprised 56 five-option multiple-choice items administered in a single 3 × 45-minute session. Prior to testing, students received standardized written and verbal instructions regarding the test purpose, answer procedures, and the requirement to work independently. The test was conducted in regular classrooms under teacher supervision, with adjusted seating to minimize collaboration and without access to textbooks or electronic devices. Before full implementation, the instrument had undergone expert validation and revision to address ambiguous wording and unclear visuals, following established content validity procedures (Haynes *et al*., 1995; A'yun *et al*., 2022). During administration, the researcher and teacher monitored the session to ensure procedural consistency without providing academic assistance (Cohen *et al*., 2018). Responses were recorded on structured answer sheets and analyzed using ANATES software for item analysis (Karnengsih *et al*., 2021; Suryanto & Taseman, 2022).

Figure 2. Trial of The Questions

Evaluation Phase

The evaluation phase involved systematic analysis of empirical data obtained from the field trial using ANATES version 4.0.9. Rather than merely reporting descriptive statistics, the evaluation focused on interpreting item performance in relation to the research objective namely, developing a psychometrically sound instrument to measure students' critical thinking skills in Biology. This evaluation was conducted by researchers by analyzing the research data obtained. A description of the evaluation results is presented in the data description and analysis. This allowed for determining which questions were suitable for use, revision, or disposal to produce a valid, reliable, and high-quality instrument. The summary of the results of the item analysis using the ANATES 4.0.9 software was as in Table 5 below.

Table 5. Results of Analysis of Trial Question Answers

Answer Analysis	Criteria	Number of Questions	Question Number
Validity	Very Valid	19	11, 12, 14, 16, 17, 18, 19, 20, 24, 25, 26, 28, 35, 39, 40, 45, 46, 51, 53
	Valid	11	1, 4, 6, 22, 27, 29, 32, 36, 37, 48, 54
	Not Valid	26	2, 3, 5, 7, 8, 9, 10, 13, 15, 21, 23, 30, 31, 33, 34, 38, 41, 42, 43, 44, 47, 49, 50, 52, 55, 56
Reliability			0,68
Difficulty Level	Very Difficult	9	8, 36, 38, 42, 47, 49, 50, 51, 55
	Difficult	18	3, 4, 5, 6, 9, 13, 15, 23, 33, 34, 37, 40, 43, 45, 52, 53, 54, 56
	Moderate	18	1, 2, 10, 12, 14, 16, 19, 21, 24, 25, 29, 31, 35, 39, 41, 44, 46, 48
	Easy	5	7, 17, 20, 27, 30
	Very easy	6	11, 18, 22, 26, 28, 32
Discrimination Index	Poor	21	2, 3, 5, 8, 9, 13, 23, 26, 28, 30, 32, 34, 36, 38, 42, 43, 44, 47, 49, 50, 55
	Satisfactory	25	1, 4, 6, 7, 10, 11, 12, 16, 17, 18, 21, 22, 25, 27, 29, 31, 33, 35, 37, 41, 48, 51, 53, 54, 56
	Good	7	14, 19, 20, 24, 40, 45, 46
	Excellent	1	39
	Not Good	2	15, 52

Based on the analysis results in Table 5, 30 questions were obtained with valid and very valid criteria and 26 questions with invalid criteria; the reliability was 0,68; the difficulty level with criteria of very easy 6 questions, easy 5 questions, moderate and difficult 18 questions, very difficult 9 questions; the discrimination index with excellent criteria 1 question, good 7 questions, satisfactory 25 questions, poor 21 questions, and not good 2 questions. The developed instrument is valid. This can be seen from the assessment of the validators in the aspects of material, construction, and language which are very good, so it is effective for measuring students' critical thinking skills. The process of testing the evaluation instrument involves construct validation by experts, followed by empirical validity testing on critical thinking skills questions.

The item validity analysis revealed that 30 out of 56 items met the criteria for valid or very valid classification, while 26 items were categorized as invalid. The relatively high

proportion of invalid items (46%) indicates that nearly half of the initial item pool did not function adequately in discriminating students' overall test performance. Further examination showed that many invalid items overlapped with those classified as having poor discrimination indices, suggesting that weak item-total correlations were associated with limited ability to differentiate between high- and low-performing students. Several of these items were also categorized as either very difficult or very easy, indicating that extreme difficulty levels may have contributed to reduced discriminatory power. This pattern suggests that certain items may not have effectively operationalized the intended critical thinking indicators, possibly requiring clearer contextual framing, improved distractor quality, or stronger alignment with students' prior knowledge.

The reliability coefficient (Cronbach's Alpha = 0.68) indicates moderate internal consistency. While this value approaches the commonly accepted threshold of 0.70 for research instruments, it suggests that the instrument, in its initial 56-item form, is more suitable for group-level measurement rather than high-stakes individual diagnostic decisions. The reliability value is likely influenced by the presence of invalid and low-discrimination items; therefore, removal or revision of problematic items is expected to improve internal consistency. This finding underscores the importance of empirical testing beyond expert judgment, as theoretical validity alone does not guarantee satisfactory psychometric performance.

The difficulty index analysis demonstrated a broad distribution of item difficulty: 6 very easy, 5 easy, 18 moderate, 18 difficult, and 9 very difficult items. Although a balanced range of difficulty is desirable to capture diverse ability levels, the concentration of items in the difficult and very difficult categories may have increased cognitive load and contributed to lower overall discrimination in some items. From a critical thinking measurement perspective, moderate-difficulty items with satisfactory or good discrimination indices are generally the most informative for distinguishing levels of reasoning ability. The discrimination index analysis further reinforced this interpretation. Only one item achieved an excellent classification and seven were categorized as good, while 21 items were classified as poor and two as not good. The concentration of weak discrimination among invalid items indicates that these questions failed to differentiate effectively between students with stronger and weaker critical thinking skills. Therefore, retention decisions were not based solely on expert validation results but on empirical performance criteria. Items meeting acceptable standards of validity and discrimination were retained, whereas invalid or poorly discriminating items were either revised or eliminated from the final instrument pool.

These findings demonstrate that expert validation and empirical validation serve complementary functions in instrument development. While expert review ensures conceptual and content alignment, empirical testing evaluates actual measurement performance . The empirical results in this study highlight areas requiring refinement and confirm that instrument development is an iterative process involving data-driven revision rather than simple confirmation of initial design assumptions. Through this evaluation process, a reduced and psychometrically improved set of items was identified for inclusion in the final version of the critical thinking assessment instrument.

The Validity

Item analysis constitutes a fundamental procedure in instrument development, as it identifies well-functioning items, detects deficiencies, and guides systematic refinement. The empirical validity analysis in this study indicated that 30 of the 56 items met the established validity criteria, whereas 26 items did not. This distribution is consequential because validity represents the core requirement of any assessment instrument ensuring that it meaningfully measures the intended construct. The substantial proportion of invalid items suggests that initial theoretical alignment with critical thinking indicators was not consistently supported by empirical performance evidence.

From a construct validity perspective, validity refers to the extent to which test items accurately represent the underlying theoretical framework. Although expert judgment initially confirmed alignment with Ennis's critical thinking indicators, empirical findings revealed that several items failed to operationalize these dimensions effectively. Many invalid items coincided with low discrimination indices and extreme difficulty levels, indicating misalignment between intended cognitive processes and actual student response patterns. Some very difficult items may have imposed excessive cognitive load or required prerequisite knowledge beyond the targeted competency, thereby weakening their construct representation.

Empirical (criterion-related) validity was examined through item–total correlation coefficients, assessing the relationship between individual item performance and overall test scores. Items with weak or non-significant correlations were categorized as invalid, suggesting that they did not function consistently with the overall critical thinking construct. Such items likely measured peripheral skills, such as factual recall, or contained ambiguous wording that distorted response consistency. As noted by Shaw and Crisp (2011), a test is valid only insofar as it measures what it claims to measure; therefore, empirical evidence must substantiate expert-based judgments.

Importantly, validity should be viewed as an iterative, evidence-based process rather than a fixed status (Elangovan & Sundaravel, 2021). The identification of 26 invalid items does not undermine the instrument as a whole; instead, it provides diagnostic insight for targeted revision. Items demonstrating satisfactory validity and discrimination were retained, while problematic items were revised or removed to enhance construct alignment and internal coherence. Consequently, the combined application of theoretical mapping and empirical analysis strengthened the instrument's capacity to operationalize critical thinking skills in Biology in accordance with established principles of educational measurement.

The Reliability

A high-quality evaluation instrument is defined not only by its validity but also by the stability and internal consistency of its scores, as well as its capacity to distinguish between different levels of student performance. Reliability analysis conducted using ANATES 4.0.9 yielded a Cronbach's Alpha coefficient of 0.68, indicating moderate internal consistency. This coefficient suggests that the instrument demonstrates acceptable score stability for exploratory, developmental, and group-level research purposes.

However, it remains slightly below the commonly recommended threshold of 0.70 for high-stakes or individual-level assessment. Accordingly, while the instrument can be considered sufficiently reliable for formative evaluation and research applications, its use In the process of making individual diagnostic or summative decisions, it must be done carefully so that errors in interpretation do not occur. As for the reliability of test questions, it can be seen in Figure 3.

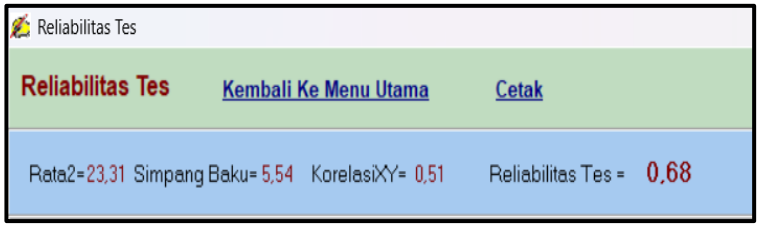

Figure 3. The Reliability of The Test Item

Figure 3. presents the reliability output generated by ANATES 4.0.9, indicating a Cronbach's Alpha coefficient of 0.68, with a mean score of 23.31, a standard deviation of 5.54, and an average item–total correlation of 0.51. This coefficient reflects moderate internal consistency, suggesting that the instrument demonstrates acceptable score stability for exploratory and developmental research, although it remains slightly below the commonly recommended 0.70 threshold for high-stakes individual assessment. Reliability represents the extent to which test scores are consistent and minimally influenced by measurement error.The moderate coefficient observed may be attributed to empirical factors identified during item analysis, including the presence of items with low validity and weak discrimination indices, which reduce item homogeneity. Additionally, the concentration of very difficult items and the extended testing duration (3 × 45 minutes) may have introduced cognitive fatigue and random error variance unrelated to the intended critical thinking construct.

Although expert validation confirmed conceptual alignment, empirical reliability evidence highlights the need for refinement to enhance consistency and replicability (Fajaryati *et al*., 2021). Data-driven improvement strategies include revising or eliminating items with low item total correlations, strengthening distractor plausibility, rebalancing item difficulty toward the moderate range, and conducting further trials with larger and more diverse samples (Badrujaman, 2020). Given that discrimination power directly influences reliability improving weak items is expected to increase internal consistency in subsequent iterations. Overall, the reliability coefficient of 0.68 indicates developmental adequacy rather than final psychometric robustness, reinforcing that instrument development is an iterative process requiring continuous empirical refinement before broader educational implementation.

The Difficulty Level

Classical Test Theory (CTT) remains a dominant framework for evaluating educational test quality, emphasizing item difficulty and discrimination as primary indicators of psychometric performance. However, because CTT parameters are sample-

dependent, item statistics must be interpreted within the empirical testing context rather than solely against theoretical benchmarks. The analysis revealed a distribution of 9 very difficult items, 18 difficult items, 18 moderate items, and 11 easy to very easy items. While the presence of 18 moderate items supports measurement precision since items within this range typically optimize score variance and discrimination, the concentration of 27 difficult and very difficult items suggests a high overall cognitive demand. Although such rigor is conceptually aligned with assessing higher-order thinking skills, an excessive proportion of highly difficult items may increase measurement error and reduce student persistence (Darling-Hammond *et al*., 2020).

More critically, the presence of 11 easy and very easy items warrants scrutiny, given that the instrument targets critical thinking, which inherently involves complex reasoning processes such as evaluation and synthesis (Grohs *et al*., 2018). Easy items may indicate that certain questions measure lower-level cognitive skills, potentially weakening construct representation particularly when accompanied by low discrimination indices. Nevertheless, a limited number of easier items can serve diagnostic and motivational functions if strategically aligned with foundational reasoning rather than recall. Overall, the empirical difficulty distribution reflects partial alignment with the intended higher-order construct but reveals imbalance. Iterative refinement is therefore necessary to recalibrate extreme items, enhance cognitive activation in easier questions, and achieve a more proportionate distribution that supports valid and reliable measurement of critical thinking skills.

The Discrimination Index

Discrimination index analysis assesses the extent to which test items differentiate between high- and low-ability students, reflecting an item's effectiveness in capturing variations in mastery of the measured construct (Odukoya *et al*., 2018). Analysis of 56 items using ANATES 4.0.9 indicated that 33 items demonstrated satisfactory to excellent discrimination, whereas 23 exhibited poor or very low indices. Weak discrimination was largely associated with extreme difficulty levels: very difficult items were answered incorrectly by most students, while very easy items were answered correctly regardless of ability, both conditions limiting score variance. Qualitative review further revealed structural weaknesses, including ambiguous wording, excessive prerequisite knowledge demands, and overemphasis on recall rather than higher-order reasoning. Consequently, item revision was guided by integrated evidence from validity, difficulty, and discrimination analyses. After iterative refinement, 22 items were retained, demonstrating balanced difficulty and acceptable discrimination, thereby strengthening internal coherence and measurement precision.

The study is grounded in the growing emphasis on cultivating higher-order thinking skills within 21st-century and competency-based education. Critical thinking encompassing interpretation, analysis, evaluation, inference, explanation, and self-regulation (Manassero-Mas *et al*., 2022) requires assessment tools capable of capturing reasoning processes rather than factual recall (Alvionita *et al*., 2020). To address the limited availability of subject-specific instruments, an initial pool of 56 multiple-choice

biology items aligned with national curriculum outcomes was developed and piloted with 103 Grade 11 students in Jambi City. Classical Test Theory analysis showed 30 items met minimum validity thresholds, with overall reliability (KR-20) of 0.68, indicating moderate internal consistency appropriate for developmental research. Multi-criteria evaluation considering validity, discrimination, difficulty, and reliability contribution resulted in a refined 22-item instrument suitable for broader implementation.

Response pattern and distractor analyses indicated that students particularly struggled with evidence-based reasoning, experimental data interpretation, and evaluation of alternative explanations, suggesting gaps in analytical integration rather than random guessing. These findings informed the design of inquiry-oriented, reasoning-focused items aimed at eliciting higher-order thinking. Given that accurate measurement informs instructional decisions, the instrument provides a psychometrically grounded tool to support the selection of effective pedagogical strategies, including inquiry-based, skills-based, and practical learning approaches shown to enhance critical thinking (Arifin *et al*., 2025). Overall, the study demonstrates that rigorous psychometric analysis combined with theoretical alignment is essential for producing defensible assessments capable of meaningfully supporting critical thinking development in biology education.

CONCLUSION

This study establishes the initial psychometric robustness of a biology critical thinking assessment developed through systematic expert validation and empirical item analysis. The instrument demonstrates strong content validity and acceptable internal consistency (r = 0.68), meeting the threshold for exploratory and formative research applications. The refinement process, which reduced 56 items to 22, enhanced discrimination power and internal coherence, although it narrowed construct coverage. Given the sample-dependent nature of Classical Test Theory parameters and the contextual scope of the study, the findings should be interpreted with caution. Practically, the instrument offers an evidence-informed prototype for formative classroom assessment of critical thinking in biology education, while methodologically emphasizing the necessity of iterative validation, construct alignment, and reliability optimization in instrument development. For broader applicability and potential high-stakes use, future research should prioritize large-scale validation, cross-context calibration, expansion of the item pool, and the application of more advanced measurement models to strengthen reliability, parameter stability, and construct representation.

REFERENCES

Abidin, Z., & Fatimatuzzaro, F. (2025). Development of a critical thinking skills assessment instrument for students on diffusion and osmosis subject. *Jurnal Mangifera Edu, 9*(2), 86-95. https://doi.org/10.31943/mangiferaedu.v9i2.216

Akbar, S. (2013). *Instrumen perangkat pembelajaran*. Bandung: PT Remaja Rosdakarya.

Almunawarah, R., Halim, A., & Elisa, E. (2023). Analysis of high school students' critical thinking skills using FRISCO indicators. *International Journal of Research and Review, 10*(10), 276-283. https://doi.org/10.21275/SR231002095937

Alvionita, D., Prayitno, B. A., & Sugiyarto. (2020). Problem-based learning with iSpring assisted inquiry learning to improve students' critical thinking skills. *Journal of Physics: Conference Series*, *1567*(4), 042044. https://doi.org/10.1088/1742-6596/1567/4/042044

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Arifin, Z., Setiawan, A., & Widodo, A. (2025). Developing critical thinking through inquiry-based learning in biology education: A meta-analysis. *Thinking Skills and Creativity*, *55*, 101678. https://doi.org/10.1016/j.tsc.2024.101678

Arikunto, S. (2015). *Dasar-dasar evaluasi pendidikan* (2nd ed.). Jakarta: Bumi Aksara.

Ariyanti, F., & Rahayu, W. P. (2025). Development of Nearpod-based learning evaluation to measure students' critical thinking skills in retail business management subjects. In *Proceedings of the 8th International Research Conference on Economic and Business (IRCEB 2024)* (pp. 62-80). Atlantis Press. https://doi.org/10.2991/978-94-6463-722-9_7

A'yun, Q., Khasanah, U., Fatmaryanti, S. D., & Sulisworo, D. (2022). Development of higher order thinking skill (HOTS) test on magnetic field concepts to improve students' critical thinking skills. *Biosfer: Jurnal Tadris Biologi*, *13*(2), 183-192. https://doi.org/10.24042/biosfer.v13i2.13658

Badrujaman, A. (2020). *Teori dan aplikasi evaluasi program bimbingan dan konseling*. Jakarta: Prenadamedia Group.

Branch, R. M. (2009). *Instructional design: The ADDIE approach*. Springer. https://doi.org/10.1007/978-0-387-09506-6

Braun, V., & Clarke, V. (2021). *Thematic analysis: A practical guide*. SAGE Publications. https://doi.org/10.1002/9781118901731.iecrm0249

Cohen, L., Manion, L., & Morrison, K. (2018). *Research methods in education* (8th ed.). Routledge. https://doi.org/10.4324/9781315456539

Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.). SAGE Publications.

Darling-Hammond, L., Flook, L., Cook-Harvey, C., Barron, B., & Osher, D. (2020). Implications for educational practice of the science of learning and development. *Applied Developmental Science*, *24*(2), 97-140. https://doi.org/10.1080/10888691.2018.1537791

DeVellis, R. F. (2017). *Scale development: Theory and applications* (4th ed.). SAGE Publications. https://doi.org/10.4135/9781506335194

Dini, N. A. I., & Kuswanto, H. (2025). Integrating local wisdom: Innovative assessment instrument of critical thinking skills in science learning. *Jurnal Eduscience*, *12*(3). https://doi.org/10.36987/jes.v12i3.6849

Elangovan, N., & Sundaravel, E. (2021). Method of preparing a document for survey instrument validation by experts. *Journal of Health and Allied Sciences*, *11*(3), 147-154. https://doi.org/10.1055/s-0041-1728878

Ennis, R. H. (1991). Critical thinking: A streamlined conception. *Teaching Philosophy*, *14*(1), 5-24. https://doi.org/10.5840/inquiryctnews201126215

Ennis, R. H. (2018). Critical thinking across the curriculum: A vision. *Topoi*, *37*(1), 165-184. https://doi.org/10.1007/s11245-016-9401-4

Erfan, M., Maulyda, M. A., Hidayati, V. R., Astria, F. P., & Ratu, T. (2020). Analisis kualitas soal kemampuan membedakan rangkaian seri dan paralel melalui tes berbasis online. *Jurnal Penelitian Pendidikan IPA*, *6*(2), 191-195. https://doi.org/10.29303/jppipa.v6i2.420

Facione, P. A. (2020). *Critical thinking: What it is and why it counts*. Insight Assessment. Retrieved from https://www.insightassessment.com

Fadlilah, A. N., & Indana, S. (2025). Development of E-LKPD based on science literacy to train critical thinking skills in Merdeka Curriculum. *Berkala Ilmiah Pendidikan Biologi (BioEdu)*, *14*(3), 553-562. https://doi.org/10.26740/bioedu.v14n3.p553-562

Fajar, N., & Suryani, R. D. (2023). Biology learning evaluation module development based on higher order thinking skills and local wisdom value. *JPBIO (Jurnal Pendidikan Biologi), 8*(1), 142-152. https://doi.org/10.31932/jpbio.v8i1.2307

Fajaryati, N., Budiyono, B., & Akhyar, M. (2021). Developing an instrument for assessing the feasibility of vocational high school students' entrepreneurial intentions. *Jurnal Pendidikan Vokasi, 11*(1), 45-56. https://doi.org/10.21831/jpv.v11i1.36789

Fereday, J., & Muir-Cochrane, E. (2006). Demonstrating rigor using thematic analysis: A hybrid approach of inductive and deductive coding and theme development. *International Journal of Qualitative Methods, 5*(1), 80-92. https://doi.org/10.1177/160940690600500107

Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2012). *How to design and evaluate research in education* (8th ed.). McGraw-Hill.

Grohs, J. R., Kirk, G. R., Soledad, M. M., & Knight, D. B. (2018). Assessing systems thinking: A tool to measure complex reasoning through ill-structured problems. *Thinking Skills and Creativity, 28*, 110-123. https://doi.org/10.1016/j.tsc.2018.03.003

Halpern, D. F. (2014). *Thought and knowledge: An introduction to critical thinking* (5th ed.). Psychology Press. https://doi.org/10.4324/9781315885279

Haynes, S. N., Richard, D. C., & Kubany, E. S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment, 7*(3), 238-247. https://doi.org/10.1037/1040-3590.7.3.238

Karnengsih, K., Harahap, R. D., & Siregar, I. H. (2021). Analisis butir soal ujian sekolah mata pelajaran IPA menggunakan program ANATES. *Jurnal Basicedu, 5*(5), 4061-4070. https://doi.org/10.31004/basicedu.v5i5.1489

Kemendikbudristek. (2022a). *Salinan Keputusan Kepala Badan Standar, Kurikulum, dan Asesmen Pendidikan Nomor 033/H/KR/2022 tentang Capaian Pembelajaran*. Jakarta: Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi. Retrieved from https://kurikulum.kemdikbud.go.id

Kemendikbudristek. (2022b). *Panduan pembelajaran dan asesmen pendidikan anak usia dini, pendidikan dasar, dan pendidikan menengah*. Jakarta: Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi. Retrieved from https://kurikulum.kemdikbud.go.id

Khoeriyah, Z., Novitasari, A., & Paratama, A. O. S. (2025). The impact of the science-technology-society model on the enhancement of student's HOTS: A systematic literature review. *Journal of Innovative Science Education, 14*(3). https://doi.org/10.15294/jise.v14i3.29625

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory Into Practice, 41*(4), 212-218. https://doi.org/10.1207/s15430421tip4104_2

Manassero-Mas, M. A., Moreno-Salvo, A., & Vázquez-Alonso, Á. (2022). Development of an instrument to assess critical thinking in science and technology. *Education Sciences, 12*(3), 201. https://doi.org/10.3390/educsci12030201

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica, 22*(3), 276-282. https://doi.org/10.11613/BM.2012.031

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749. https://doi.org/10.1037/0003-066X.50.9.741

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill.

O'Connor, C., & Joffe, H. (2020). Intercoder reliability in qualitative research: Debates and practical guidelines. *International Journal of Qualitative Methods, 19*, 1-13. https://doi.org/10.1177/1744987120927206

Odukoya, J. A., Adekeye, O., & Okonkwo, E. (2018). Assessing the effectiveness of mobile learning devices in tertiary institutions: The experience of undergraduates in a private university in Nigeria. *Cogent Education, 5*(1), 1540918. https://doi.org/10.1080/2331186X.2018.1540918

Pamelia, S. S., & Hariani, D. (2021). Analysis of critical thinking ability of students class X SMA Negeri 1 Sampang on environmental pollution material. *Berkala Ilmiah Pendidikan Biologi (BioEdu), 11*(1), 107-115. https://doi.org/10.26740/bioedu.v11n1.p107-115

Partnership for 21st Century Learning. (2019). *Framework for 21st century learning definitions*. Battelle for Kids. Retrieved from https://www.battelleforkids.org

Paul, R. (2018). Critical thinking and the critical person. In *Thinking: The second international conference* (pp. 373-404). Routledge. https://doi.org/10.4324/9781315802015-27

Rahmi, Y. L., Miatidini, N. A., Alberida, H., Darussyamsyu, R., Ichsan, I. Z., Sigit, D. V., Titin, T., Koc, I., & Sison, M. H. (2021). HOTS assessment of biology cell: Validity, practicality and reliability. *Jurnal Penelitian Pendidikan IPA, 7*(3), 481-487. https://doi.org/10.29303/jppipa.v7i3.742

Ratna, R., Suryanda, E., & Rusdi, R. (2025). Development of critical thinking skills assessment instrument based on Ennis framework on environmental change material. *Biosfer: Jurnal Pendidikan Biologi, 18*(1), 1-12. https://doi.org/10.21009/biosferjpb.54197

Retnawati, H. (2016). *Validitas, reliabilitas, dan karakteristik butir: Panduan untuk peneliti, mahasiswa, dan psikometrian*. Yogyakarta: Parama Publishing. Retrieved from https://staffnew.uny.ac.id/upload/132255129/pengabdian/buku-validitas-reliabilitas.pdf

Santosa, T. A., Lufri, L., & Andromeda, A. (2023). Development of higher order thinking skills (HOTS) instruments in biology learning: A systematic review. *Jurnal Mangifera Edu, 8*(1), 1-14. https://doi.org/10.31943/mangiferaedu.v8i1.166

Sari, D. N., Sunarno, W., & Prayitno, B. A. (2024). Diagnostic assessment profile of learning styles and critical thinking skills in biology learning based on the Merdeka Curriculum. *JPBI (Jurnal Pendidikan Biologi Indonesia), 10*(3), 876-887. https://doi.org/10.22219/jpbi.v10i3.36557

Schwab, K. (2017). *The fourth industrial revolution*. Currency.

Septiany, L. D., Puspitawati, R. P., Susantini, E., Budiyanto, M., Purnomo, T., & Hariyono, E. (2024). Analysis of high school students' critical thinking skills profile according to Ennis indicators. *IJORER: International Journal of Recent Educational Research, 5*(1), 157-167. https://doi.org/10.46245/ijorer.v5i1.544

Shaw, S., & Crisp, V. (2011). Tracing the evolution of validity in educational measurement. *Assessment in Education: Principles, Policy & Practice, 18*(4), 365-382. https://doi.org/10.1080/0969594X.2011.607444

Sintia, D. N., & Yuliani, Y. (2024). Analysis of students' critical thinking skills in biology learning at senior high school. *Berkala Ilmiah Pendidikan Biologi (BioEdu), 13*(2), 334-343. https://doi.org/10.26740/bioedu.v13n2.p334-343

Sugiyono. (2019). *Metode penelitian pendidikan: Pendekatan kuantitatif, kualitatif, dan R&D*. Bandung: Alfabeta.

Suraida, S., Aslamiah, A., & Suriansyah, A. (2025). Development of assessment instruments to measure students' critical thinking skills. *International Journal of Social Science and Human Research, 8*(2), 1123-1132. https://doi.org/10.47191/ijsshr/v8-i2-48

Suryanto, S., & Taseman, T. (2022). Analisis kualitas butir soal ujian akhir semester genap mata pelajaran matematika menggunakan program ANATES. *Jurnal Basicedu, 6*(4), 7310-7320. https://doi.org/10.31004/basicedu.v6i4.3321

Syafril, S., Aini, N. R., Pahrudin, A., & Yaumas, N. E. (2021). Developing instrument for students' critical thinking ability on mathematics. *European Journal of Educational Research, 10*(1), 337-349. https://doi.org/10.12973/eu-jer.10.1.337

Ulfa, M., & Kuswanti, N. (2020). Development of assessment instrument based on higher order thinking skills of respiratory system of grade XI of senior high school. *Berkala Ilmiah Pendidikan Biologi (BioEdu), 9*(3), 431-437. Retrieved from https://ejournal.unesa.ac.id/index.php/bioedu/article/view/35726

Wahyuningtyas, A., Triwahyuni, E., & Kustiyowati. (2025). Learning media based on Google Sites to improve critical thinking in senior high school students. *Jurnal Pedagogi dan Pembelajaran, 8*(2), 289-301. https://doi.org/10.23887/jp2.v8i2.99322

Willingham, D. T. (2019). *Why don't students like school?: A cognitive scientist answers questions about how the mind works and what it means for the classroom* (2nd ed.). Jossey-Bass. Retrieved from https://www.wiley.com

World Economic Forum. (2020). *Schools of the future: Defining new models of education for the fourth industrial revolution*. WEF. Retrieved from https://www.weforum.org

Yokhebed, Y., Karmadi, R. M. D., & Nastiti, L. R. (2025). Validity and reliability analysis of a socioscientific issues-based critical thinking self-assessment instrument using the Rasch model. *JPBI (Jurnal Pendidikan Biologi Indonesia), 11*(1), 73-82. https://doi.org/10.22219/jpbi.v11i1.38902

Yu, P. L. H., & Zin, Z. M. (2023). The development of higher order thinking skills (HOTS) assessment instrument for biology education: A systematic literature review. *Asian Journal of University Education, 19*(4), 789-802. https://doi.org/10.24191/ajue.v19i4.24567